# 単語間類似度を用いた文の接続性の定量化

情報科学科　前山　隼大　　指導教員：山村　毅

## 1　はじめに

自然言語は特筆すべき専門知識を必要とせず、人間が平等に使用できるツールである。今日、自然言語を用いて行う機械と人間との対話、その中でも自然な対話 (人間にとって違和感のない対話) に注目が集まっている。人間と人間との対話は、相手が発した文に対して適切な応答を互いに繰り返すことで行われるが、機械と人間との間にこれを実現させるためには文と文の接続性、いわゆる文の繋がりの違和感のなさを定量化する必要がある。

本研究では、文の接続性の問題を「ある文の次に来る文の確率」という観点から捉え、TOEIC の応答問題を対象に、基礎単語及び単語間類似度を用いて、質問文に対する適切な応答文を選択する手法の開発を行う。

## 2　先行研究

### 2.1　概要

著者ら [1] は、応答問題の質問文を $q = (q_1, q_2, \cdots, q_n)$、応答文を $r = (r_1, r_2, \cdots, r_m)$ としたとき ($q_i, r_i$ はそれぞれ質問文・応答文の特徴)、以下の評価関数 (1) を最大化するものを適切な応答文として選択する手法を提案した。

$$f(q, r) = \prod_{i=1}^{m} P(r_i) \prod_{j=1}^{n} P(q_j | r_i) \tag{1}$$

### 2.2　精度と問題点

質問文 1 文に対して 3 つの応答文の選択肢がある 1038 例について、47 個の特徴を用いて応答文の選択実験を行った結果、最高で 53.7% の精度 (判定不能文数 67) を得た。しかし、以下に挙げる複数の問題点を抱えている。

- 特徴にヒューリスティックなものを使用している
  用いた特徴は著者ら自身がヒューリスティックに選定したものであるので、全ての文を特徴表現するのに十分でない。
- 精度が低い
  精度 53.7% はシステムとして実装するには精度が低い。
- 判定不能文数が多い
  パターン分類を行う際に応答文選択肢に特徴が 1 つも存在しない場合など、判定不能な文が多く存在する。

これらの問題点について解決する必要がある。

## 3　基礎単語と単語間類似度

### 3.1　基礎単語

基礎単語[*1]は Charles Kay Ogden によって定義された、日常生活を行う上で必要十分な単語として選定された英単語 850 語である。基礎単語として挙げられていない単語は、基礎単語の組み合わせや言い換えにより表現できる。

基礎単語が単語の中でも日常会話で使われる単語を中心に抽出されていることから、この基礎単語を特徴として用いることで、日常で使用される多くの文のパターン分類に利用できると考えられる。また、明確に定義されているため、ヒューリスティックな特徴を使用せずに済む。

### 3.2　単語間類似度

パターン分類を行う際、文中に特徴が全く存在しない場合は判定不能となる。そこで単語間類似度を利用し、特徴以外の単語を特徴である基礎単語へ変換し、特徴情報を増やすことで判定不能の問題に対処する。この際、2 単語間がどれほど似ているかを表す単語間類似度を用いる。

単語間類似度の算出には WordNet[*2]のシソーラスを用いる。WordNet では単語の意味が階層木構造で構築されており、各ノード間の距離 $d$ を用いて $\frac{1}{d+1}$ を単語間類似度として定義している。

## 4　実験と結果

### 4.1　実験方法

先行研究と同じデータである、質問文 1 文に対して 3 つの応答文の選択肢がある 1038 例について、基礎単語以外の単語を単語間類似度 (閾値 $DOS = \frac{1}{d+1}$) を用いて基礎単語へ変換し、基礎単語を特徴として応答文の選択実験を行った。閾値 $DOS$ は自由パラメータ $d$ について 0.0 から 8.0 まで 0.5 刻みで変えて、正解率を求めた。分類器にはナイーブベイズ分類器、評価には 10 分割交差検定を用い、ゼロ頻度問題の対応には加算スムージングを用いた。

### 4.2　実験結果

$d$ を大きくするほど高い正解率となったが、$d \geq 5.5$ は横ばいであった。最大正解率のときの結果を以下の表 1 に示す。

表 1　実験結果

| 正解 | 不正解 | 判定不能 | 全例題数 | 正解率 (%) |
|---|---|---|---|---|
| 700 | 324 | 14 | 1038 | 68.4 |

この結果は、正解率について先行研究 [1] に比べ、適合度検定 (有意水準 1%) で有意な差がある。また、判定不能文数も 14 と先行研究に比べ大幅に減少している。

## 5　まとめ

質問文 1 文に対し応答文 3 文の選択肢があるデータに対し、単語間類似度を用いた基礎単語変換を施し、特徴である基礎単語の抽出とその特徴を基にした文の接続性の定量化を行った。ナイーブベイズ分類器により実験を行った結果、最も高い正解率として 68.4% を得ることができ、先行研究との有意差を確認することができた。今後の課題としては、基礎単語のカテゴリ分類による精度向上や、判定不能文数を 0 にすることなどが挙げられる。

## 参考文献

[1] 前山, 秋田, 山村: ”応答問題を用いた文の接続性の定量化”, 電気・電子・情報関係学会東海支部連合大会講演論文集, L3-6, 2014

[*1] Ogden's Basic English - http://ogden.basic-english.org/

[*2] About WordNet - http://wordnet.princeton.edu/